

# INTERACTIONS, NEIGHBORHOOD SELECTION AND HOUSING DEMAND<sup>1</sup>

Yannis M. Ioannides and Jeffrey E. Zabel

Department of Economics, Tufts University, Medford, MA 02155

May 17, 2002

## Abstract

This paper contributes to the growing literature that identifies and measures the impact of social context on individual economic behavior. We develop a model of housing demand with neighborhood effects and neighborhood choice. Modelling neighborhood choice is of fundamental importance in estimating and understanding endogenous and exogenous neighborhood effects. That is, to obtain unbiased estimates of neighborhood effects, it is necessary to control for non-random sorting into neighborhoods.

Estimation of the model exploits a unique data set of household data that has been augmented with contextual information at two different levels (“scales”) of aggregation. One is at the neighborhood cluster level, of about ten neighbors, with the data coming from a special sample of the American Housing Survey. A second level is the census tract to which these dwelling units belong. Tract-level data are available in the Summary Tape Files of the decennial Census data. We merge these two data sets by gaining access to confidential data of the U.S. Bureau of the Census. We overcome some limitations of these data by implementing some significant methodological advances in estimating discrete choice models.

Our results for the neighborhood choice model indicate that individuals prefer to live near others like themselves. This can perpetuate income inequality since those with the best opportunities at economic success will cluster together. The results for the housing demand equation are similar to those in our earlier work [ Ioannides and Zabel (2000) ] where we find evidence of significant endogenous and contextual neighborhood effects.

Int-neigh-dem-514.tex

---

<sup>1</sup>The research in this paper was conducted while the authors were U.S. Census Bureau Research Associates at the Boston Research Data Center. Research results and conclusions expressed are those of the authors and do not necessarily indicate concurrence by the Bureau of the Census. This paper has been screened to ensure that no confidential data are revealed. It should not be quoted without approval from the authors. Ioannides thanks the National Science Foundation and the MacArthur Research Network for Social Interactions and Economic Inequality for generous research support. Both authors thank Jim Davis for his help with the disclosure process and Dennis Epple and Anna Hardman for generous comments and editorial suggestions. Earlier versions of the paper were presented at the MacArthur Network for Social Interactions and Economic Inequality, the “Neighborhoods and Outcomes” session of the 2002 AEA Meetings, at the University of Florida and the University of Texas. We thank Steven Durlauf for his insightful discussion in Atlanta.

# 1 Introduction

U.S. income inequality in the 1990s has led, among other reasons, to a new concern with measuring the impact of social context on economic behavior, often referred to as “social effects.” One important component of social effects is the impact of one’s place of residence, or neighborhood effects. Desirable social interactions and beneficial local community “social capital” are known to be features of neighborhood or residential communities. The rich may keep getting richer because they benefit from a better social environment than do poorer individuals.

Decisions about whether or not to move, where to locate and how much housing to consume will be influenced generally by a perception of the behavior and characteristics of current and potential neighbors. The link between place of residence and economic outcomes as an explanation for the persistence (and even increase) of income inequality in the U.S. is a popular area of both theoretical and empirical research. Yet, we still do not understand fully the decision of where to live and the factors households typically take into consideration in making this decision.

The results of this research have important policy implications. For example, affordable housing policy has evolved from the time when huge complexes were built in poor parts of U.S. cities to a contemporary interest in such policies as housing vouchers and the requirement that, in many affluent towns and cities, a certain percentage of the total housing stock must be affordable housing. This evolution of policy thinking is consistent with the idea that neighborhood effects are important determinants of economic wellbeing. The affordable housing experiment that is known as “Moving To Opportunity” [ Katz *et al.* (2000) ] is a specific example of a program that attempts to measure the impact of households that move from poor neighborhoods to more economically advantaged areas.

In this paper, we emphasize the role of neighborhood effects in influencing households’ residential location decisions and how such decisions affect, in turn, neighborhood effects. We extend our earlier work [ Ioannides and Zabel (2000) ], where we estimate a model of housing demand with neighborhood interactions, to allow for neighborhood choice. We estimate this new model by utilizing a unique household-level data set that has been augmented with contextual information at two different levels (“scales”) of aggregation. One is at the neighborhood cluster level, of about ten neighbors, with data from a special sample of the American Housing Survey (AHS). A second

level is the census tract, to which the dwelling units in the above sample actually belong. Tract-level data are available in the Summary Tape Files (STF3) of the decennial Census data. We merge these two data sets by gaining access to confidential data of the U.S. Bureau of the Census.

Our utilization of micro data allows us to implement recent methodological advances in the econometrics of interactions-based models [ Brock and Durlauf (2001) ]. Our approach sets out to complement several other studies, notably the recent papers by Epple and Sieg (1999) and Epple, Romer, and Sieg (2001), who use community-level data for the Boston MSA in 1980. Nesheim (2001) uses micro data from the National Educational Longitudinal Study of 1988 and treats the choice of neighborhood as a continuous decision identified by the mean education in one's neighborhood.

Ioannides and Zabel (2000) find strong evidence of both endogenous neighborhood effects, that is of individuals' caring about the decisions of their neighbors, and of exogenous neighborhood effects, that is of individuals' caring about the socioeconomic characteristics of their neighbors. We study housing demand by treating members of small neighborhoods, to be referred as neighborhood clusters, as groups of interacting agents. Under the Nash equilibrium assumption, each individual takes neighbors' decisions as given. Individuals who have chosen to reside in the same neighborhood cluster are likely to have correlated observable and unobservable characteristics. Unless one understands how individuals choose where to locate, little can be known about the distribution of unobservable factors across neighborhoods. Some of these factors may be important in human capital formation and responsible for the creation and maintenance of social capital. Such factors imply dependence among neighbors' housing consumption decisions and among their unobservable individual characteristics.

In the present paper we extend our earlier work to account for neighborhood choice and housing demand as joint decisions in the presence of neighborhood effects. We model the choice of neighborhood by identifying it with the choice of community as represented by a *census tract*. Census tracts constitute geographical areas of about 5000 inhabitants that are defined so as to be relatively homogeneous in regard to the characteristics of their population. Our modelling decision is influenced by a whole host of factors, including the fact that information about neighborhood clusters is not generally available outside the particular sample of the AHS that we utilize, while descriptive information is widely available about census tracts from the U.S. Census. Conditional on location

in a census tract, individuals choose how much housing to consume. In doing so they recognize that they would be in close proximity to a small number of other households, whose characteristics and decisions may affect their own choices. It is in this sense that small groups of nearest neighbors are treated as groups of interacting agents.

We start by describing the data in Section 2 of the paper. In Section 4 we discuss additional aspects of the data, including the construction of the house price index and the estimation of permanent income for households, both of which are required by the estimation of the housing demand model. We develop our model of neighborhood choice and housing demand in Section 3. We also address there some important methodological issues that arise because of the limitations of our data set. In particular, since we do not observe samples of clusters in each census tract, except for those that we use to estimate the demand model, we cannot estimate the choice of neighborhood cluster conditional on tract choice. This means we cannot estimate directly an important component of the neighborhood choice model, that is the inclusive value in the second stage of the census tract choice model. Therefore, we utilize information in the data on the distributions of variables within tracts to develop proxies for the inclusive value that derive naturally from the behavioral model that are consistent with the methodological framework in McFadden (1978). Also, we only observe where individuals actually reside and not alternative tract choices. We rely on a result by McFadden (1978) that shows that consistent estimates can be obtained, when alternative tract choices are not observed, by randomly choosing from the full set of census tracts in the metropolitan area where a household resides. We also lay out the econometric methodology that allows us to obtain consistent estimates of the housing demand equation conditional on tract choice. This results in the inclusion of eleven (one for each tract in the choice set) sample selection correction terms in the housing demand equation. We implement a new result in Brock and Durlauf (2001) that provides valid instruments for identifying the endogenous neighborhood effect. In sum, our estimates of neighborhood effects utilize information on households' choices of neighborhoods and of the quantity of housing and rests on methodology that is firmly grounded in the modern theories of interactive discrete choice and of interdependent preferences.

We present the estimation results in Section 4. Our results for the neighborhood choice model show that individuals prefer to live near others like themselves. This can perpetuate income inequal-

ity since those with the best opportunities at economic success will cluster together. The results for the housing demand equation are similar to those in our earlier paper [ Ioannides and Zabel (2000) ], where we find evidence of significant endogenous and contextual neighborhood effects.

## 2 The AHS Data

We start with a discussion of the data because its special features motivate our theoretical approach and our estimation procedure. The main data source used for this study is the national version of the American Housing Survey (NAHS). The NAHS is an unbalanced panel of more than 50,000 housing units that are interviewed every two years. It serves as the basis for most U.S. housing statistics. The NAHS contains detailed information on dwelling units and their occupants through time. This information includes the current owner’s evaluation of the unit’s market value, structural characteristics of dwelling units, and self-reported information on the house’s current occupants.

In 1985, 630 dwelling units, which will be referred to as *kernels*, were selected at random, and up to ten nearest housing units, to be referred to together with the kernel as neighborhood *clusters*, were interviewed. Additional observations from larger clusters were added, making the total number of clusters equal to 680, yielding a data set of 7,350 housing units. This procedure was repeated in 1989, when 769 kernel units were selected, and in 1993, when 1018 kernel units were selected. Additional units in existing clusters were included in 1989 to reflect additional units that had been added within the perimeter of the 1985 “neighborhood.” By 1993, a maximum of 20 neighboring units were allowed per cluster.<sup>2</sup> The result is an unbalanced three-wave panel of dwelling units. Tables 1–5 in Ioannides and Zabel (2000) provide extensive details on the structure of the data, including observation counts on new clusters, new households and new units, etc., and their geographic distribution.<sup>3</sup>

By working with this special subsample of the NAHS, we generate a data set that includes information on the value and characteristics for all dwelling units in the cluster. The owner-

---

<sup>2</sup>We are grateful to Barbara T. Williams, U.S. Bureau of the Census, for this clarification.

<sup>3</sup>How good is this definition of a neighborhood in the form of clusters of dwelling units, as employed by the AHS? It is obviously appropriate for apartment buildings in densely developed US neighborhoods. Support for suburban communities comes from Gans (1967), a classic study of suburban social structure:

Neighboring rarely extended more than three or four houses away in each direction, so that the functional neighborhood usually consisted of about *ten to twelve* houses at the most, although people did say hello to everyone on the block . Gans, *op. cit.*, p. 156, as quoted by de Souza Briggs (1997), p. 211 ].

occupant’s characteristics that we work with include, in particular, the owner’s years of schooling, whether the owner is white, whether the owner is married, the number of persons in the household, household income, and whether the house has changed hands in the last five years. For each survey, owners are asked to estimate how much their property (and, in addition, its lot, if appropriate) would sell for if it were for sale. Goodman and Ittner (1992) and Kiel and Zabel (1999) find that while, on average, owners over-estimate their value by 5%, this bias is not systematically related to the observed characteristics of the owner, house, or neighborhood.

Since there typically are few clusters in a given MSA, we need to base our model of neighborhood selection on the choice of census tracts. Access to the confidential version of the NAHS provides information that allows us to identify the census tract in which each dwelling unit in the public version of the AHS is located. Census tract characteristics are available in the Summary Tape Files (STF3) for the decennial Censuses. They include demographic information such as the median household income, structural characteristics such as the median number of bedrooms, mobility information such as the percent of households that moved in the last five years, and tenure and vacancy statistics. There is also information of the distribution of a number of these variables. This is crucial for our estimation procedure, as we discuss further below. We merge the information from the 1980 and 1990 STF3s with the AHS data by census tract. We interpolate and extrapolate from the reported averages of the 1980 and 1990 STF3 data to create the tract variables for the 1985, 1989, and 1993 surveys.

An observation on a dwelling unit from the AHS in 1985, 1989, and 1993 is included in our sample only if it: is associated with a regular occupied interview, is owner occupied, lies in a metropolitan statistical area (MSA), is valued by the owner at least \$10,000, and is not missing any information on unit, occupant, or census tract characteristics that are included in our analysis. Since we are using the information about the neighbors to measure neighborhood effects, we require that there be at least four other dwelling units in the cluster after the above selection criteria have been employed. This reduces the number of observations for 1985, 1989, and 1993 to, respectively, 1747, 1954, and 2671. The 764 neighborhood clusters included in our analysis are the result of pooling over the three waves of the data. There are, on average, 8.3 observations per cluster, located in about 100 MSAs. The full frequency distribution of cluster sizes is given in Table 2.

Additional information on the distribution of clusters across the three years is given in Tables 1–5, Ioannides and Zabel (2000). The names and summary statistics for the key variables used in the present study are given in Table 1, for the AHS data, and in Table 3, for the census tract data. Table 3 compares descriptive statistics of the key variables for the census tracts where households are observed to reside with those for the ten randomly chosen tracts.

### 3 Choice of Neighborhood and Housing as Joint Decisions

Choice of neighborhood is an important decision that reflects many considerations, of which housing is a key element. Some neighborhood amenities are truly exogenous factors. Others, however, such as the ambience of a neighborhood, not only are endogenous but are hard to measure or may be truly unobservable. They may, in part, reflect who chooses to reside in a particular neighborhood and therefore will be correlated with the socioeconomic characteristics of neighbors. Since neighborhood choice is not random, it is important to know what are the key determinants of individuals' neighborhood selection. In addition, since locational quality as evaluated by households is likely to reflect their observed and unobserved characteristics, accounting for neighborhood selection is important in order to obtain unbiased estimates of the parameters in the housing demand equation, especially when it includes neighborhood effects as regressors. The present paper thus models neighborhood choice and housing demand as joint decisions, where neighborhood choice is discrete and housing demand is continuous. This is in contrast to Nesheim (2001), who uses a continuous notion of neighborhood, and to Ioannides and Zabel (2000), who take neighborhood choice as given and estimate a model of housing demand with neighborhood effects.

Our approach is designed to utilize available public data that have been augmented with confidential U.S. Census data. The public NAHS data contains information on dwelling units and their inhabitants and also allows us to identify those that belong to small groups of about ten nearest neighbors, the members of neighborhood clusters in the terminology of this paper. We identify the discrete choice of neighborhood via the choice of census tract within a particular metropolitan area, both of which are observable in our augmented data.

### 3.1 The Preference Structure

We assume that a household  $h$  that is considering occupying a dwelling unit that belongs to neighborhood cluster  $k$ ,  $k = 1, \dots, N_s$ , in tract  $s$ ,  $s = 1, \dots, S$ , would enjoy utility  $\Omega_{skh}$ , that is made up of two additive components:

$$\Omega_{skh} = V_{skh} + \epsilon_{skh}. \quad (1)$$

The first component,  $V_{skh}$ , is a *conditional* indirect utility function, that is specified below as a function of price and income, and of additional observable and unobservable characteristics of individuals residing in neighborhood cluster  $k$  and of the census tract  $s$  in which  $k$  lies. The second,  $\epsilon_{skh}$ , is a random component of utility, drawn from a distribution that is specified below, that affects neighborhood choice and is assumed to be observable by the individual and unobservable by the econometrician. This model adopts some of the features of Dubin and McFadden (1984).

We specify the conditional indirect utility function in (1) above as

$$\begin{aligned} V_{skh} &\equiv V(g_s, p_s, I_h; z_h; y_{n(h)}, z_{n(h)}; v_k + \eta_{skh}) = \omega_{skh} + [v_k + \eta_h] e^{-\delta p_s} \\ &\equiv \zeta_h g_s + \left[ \alpha + \frac{\mu}{\delta} + \mu p_s + \xi z_h + \delta I_h + \beta \Pi_y(y_{n(h)}) + \gamma \Pi_z(z_{n(h)}) \right] e^{-\delta p_s} + [v_k + \eta_h] e^{-\delta p_s}. \end{aligned} \quad (2)$$

In this definition of  $V_{skh}$ ,<sup>4</sup>  $\zeta_h, \alpha, \mu, \delta, \xi, \beta$ , and  $\gamma$  are preference parameters. Component  $\omega_{skh}$  is it-

---

<sup>4</sup>We may rationalize the emergence of such preferences as those represented in Equ. (3) as follows. Individual  $h$  cares about the quantity of housing services indirectly produced by characteristics of her property  $i$  in tract  $s$ ,  $q_i$ , and by attributes of the neighborhood where it is located,  $x_s$ . Let direct utility depend on composite nonhousing consumption,  $c_h$ , and on the consumption of housing services, which are indirectly produced by  $(q_i x_s)$ , and on own demographic characteristics that might affect preferences,  $z_h$ , and on the vectors of housing consumptions in the neighborhood,  $y_{n(h)}$  [ Pollak (1976) ], and of (observable) demographic characteristics of neighbors,  $z_{n(h)}$ ,

$$U_h = U(c_h, q_i, x_s; z_h; y_{n(h)}, z_{n(h)}).$$

Assuming a static setting an individual chooses how to allocate her income  $I_h$  to nonhousing consumption,  $c_h$ , and housing consumption indirectly through choice of  $(q_i, x_s)$ , subject to

$$c_h + P(q_i, x_s; y_{n(h)}, z_{n(h)}) = I_h,$$

where nonhousing consumption is the numeraire,  $P(q_i, x_s; y_{n(h)}, z_{n(h)})$  denotes the hedonic price function as a function of the corresponding vector of characteristics,  $(q_i, x_s)$ . Neighbors' consumption and socioeconomic characteristics,  $(y_{n(h)}, z_{n(h)})$ , are taken as given (that is, given neighborhood choice) and only  $(q_i, x_s)$  are subject to individual choice.

Given patterns of development in U.S. cities, one must account for the fact that  $(q_i, x_s)$  are not chosen independently of  $(y_{n(h)}, z_{n(h)})$ . In other words, certain combinations of structural characteristics and neighborhood attributes may typically be found in neighborhoods inhabited by households with certain socioeconomic characteristics. Nonetheless, from the viewpoint of an individual consumer it is appropriate to think of a choice set defined in terms of bundles of goods that are freely varied by individual choice, such as interior structural characteristics of dwelling units, and neighborhood socioeconomic characteristics and decisions, which are treated as quasi-fixed by individuals.

Such a "mixed" indirect utility function is akin to those used in public economics to express externalities and in consumption theory to express quasi-fixed components of the consumption bundle.



self defined as the maximum value of a direct utility function, subject to a budget constraint, with respect to housing and nonhousing consumption. The conditional utility function is thus specified as a separable function of tract-specific characteristics that are independent of the characteristics of housing choice,  $g_s$ , and of price,  $p_s$ , individual income,  $I_h$ , and a vector of own demographic characteristics that affect preferences,  $z_h$ . The conditional utility function is also affected by neighbors' decisions and characteristics, *neighborhood effects*, on preferences by being specified as conditional on general functions,  $\Pi_y(\cdot)$  and  $\Pi_z(\cdot)$ , of the vectors of one's neighbors' housing consumption,  $y_{n(h)}$ , and of neighbors' socioeconomic characteristics,  $z_{n(h)}$ . We define housing consumption to be the continuous flow of services that comes from the dwelling structure and the neighborhood amenities associated with the dwelling's location.

The terms  $v_k$  and  $\eta_h$  on the right hand side of (2) are random variables that are assumed to be unobservable by the analyst but observable by individuals. The former is a specific characteristic of neighborhood cluster  $k$ , that is considered as independent and identically distributed across neighborhood clusters within each census tract but assumed to be observed by all those households that evaluate cluster  $k$  and is thus common among them; the latter is an individual household characteristic that is independently and identically distributed over all individuals. As we discuss further below,  $v_k$  will stand for a cluster-specific random effect and  $\eta_h$  an i.i.d. stochastic shock that make up the error component of the housing demand equation.

Note that the tract-specific term  $\zeta_h g_s$  is specified as individual-specific. That is, we allow preferences for tract characteristics to vary across individuals. This allows, for example, a household with children to put a different weight on school quality than a household with no children, or a household of a given ethnic background or race to value the presence of neighbors of the same ethnic background or race differently than those from different ethnic backgrounds or races.

Using Roy's identity,  $y = -\frac{\partial \Omega}{\partial p} / \frac{\partial \Omega}{\partial I}$ , with the utility function (1) yields the housing demand function for household  $h$  conditional on census tract  $s$  and neighborhood cluster  $k$ :

$$y_{skh} = \alpha + \mu p_s + \xi z_h + \delta I_h + \beta \Pi_y \left( y_{n(h)} \right) + \gamma \Pi_z \left( z_{n(h)} \right) + v_k + \eta_h. \quad (3)$$

Notice that differentiation of  $\Omega_{skh}$  with respect to  $p_s$  and  $I_h$ , as required by Roy's identity, eliminates the terms  $\zeta_h g_s$  and  $\epsilon_{skh}$ , that is, the tract-specific observable and unobservable components of utility. Those components will play a role in the neighborhood choice stage of the problem. Below, we

specify the stochastic structure in further detail.

Invoking the terminology of Manski (1993), we say that the term  $\beta\Pi_y(y_{n(h)})$  on the RHS of (3) expresses an *endogenous social effect*: a person’s behavior depends on the actual *behavior* of her neighbors. Such a social effect is, of course, central to the notion of neighborhood effects. The term  $\gamma\Pi_z(z_{n(h)})$  expresses a *contextual effect*, an exogenous social effect which reflects taste over the characteristics of one’s neighbors such as their race, ethnicity, and income.<sup>5</sup> The unobserved stochastic components on the RHS of (3) may reflect a conditional version of what Manski calls a *correlated effect*: similar individuals are likely to make similar choices of dwelling units, of structural characteristics, and of neighborhoods and therefore have unobserved characteristics in common. Such dependence follows as an outcome of the sorting features of the matching process of households with dwelling units, whereby individuals’ interest in the socioeconomic profile of their neighborhoods is mediated through the residential matching process. So, correlated effects may express unobserved characteristics of the neighborhood, as well.

Imposing neighborhood Nash equilibrium within the neighborhood cluster requires considering a simultaneous equations system along the lines of (3), with an equation for each of the members of neighborhood cluster  $k$ . The simultaneity is evident by the presence of  $y_{n(h)}$ , the vector consisting of housing demands by an individual’s neighbors, all of which are endogenous variables, on the RHS of (3).

In the context of housing demand, the endogenous effect expresses a notion of ”keeping up with the Joneses,” whereby individuals view their neighbors’ decisions about maintenance, repair, renovation, and additions and will strive to keep up by making similar decisions and in accordance with their own preferences. The contextual effect arises as a matter of taste or when owners view their neighbors’ characteristics, e.g. income, as a signal of their future housing consumption and thus alter their own housing consumption accordingly.

### 3.2 Neighborhood Choice

Specification of the stochastic structure of  $\epsilon_{skh}$  in Equ. (1) allows us to model the choice of neighborhood as a discrete choice [McFadden (1978)]. The probability that individual  $h$  chooses

---

<sup>5</sup>We note that even if the term  $\gamma\Pi_z(z_{n(h)})$  might not be assumed as a direct contextual effect, it would show up in a reduced-form, as a result of the presence of the endogenous social effect in the structural form.

tract  $s_h$ , from among  $s = 1, \dots, S$ , and neighborhood cluster  $k_h$ , from among  $k = 1, \dots, N_s$ , is given by the probability that actual utility from this choice exceeds utility from all other choices:

$$Prob_{s_h k_h h} = Prob \left\{ \omega_{s_h k_h h} - \omega_{skh} \geq -(\epsilon_{s_h k_h h} - \epsilon_{skh}) - (v_{k_h} - v_k) e^{-\delta p_s}; \quad \forall (s, k) \neq (s_h, k_h) \right\}. \quad (4)$$

We note that, according to (3), the term  $e^{-\delta p_s}$  is common across all census tracts an individual considers because the price  $p_s$  is MSA-specific. Therefore, the term  $\eta_h e^{-\delta p_s}$  cancels out for all comparisons of alternative clusters because they all belong to the same MSA. The term  $(v_{k_h} - v_k) e^{-\delta p_s}$ , however, does not cancel out. Therefore, neighborhood choice is determined by and introduces dependence among observations for individuals who choose to locate in the same neighborhood cluster. This not accounted for directly in the neighborhood choice model. The interdependence of neighbors' demands, however, does receive a lot of attention below.

Under the assumption that the  $\epsilon_{skh}$ 's are independently and identically extreme-value distributed across all census tracts and all neighborhood clusters within them, the choice probabilities are given by the multinomial logit model (MNL):

$$P_{s_h k_h h} = \frac{e^{\omega_{s_h k_h h}}}{\sum_{s=1}^S \sum_{k=1}^{N_s} e^{\omega_{skh}}}. \quad (5)$$

A problem with applying the MNL model is that  $\epsilon_{skh}$  is unlikely to be independent across alternative residential choices. In particular, alternative cluster choices within the same census tract will include common tract-level unobservables that will cause their error terms to be correlated.

Define the set  $C_s$ ,  $s = 1, \dots, S$  to be the set of clusters in census tract  $s$  and let  $N_s = |C_s|$ . One can think of a hierarchical structure where the  $C_s$ 's constitute one level and the census tracts another. At the lower level, that is conditional on a tract, characteristics of clusters within a tract that are reflected in the  $\epsilon_{skh}$ 's are assumed to be correlated. At the higher level, characteristics of clusters across census tracts are assumed to be uncorrelated. Also, conditional on the choice of a census tract, the cluster choices are assumed to be independent.<sup>6</sup> Such a structure is the basis for the nested multinomial logit (NMNL) model.

Under the assumption of a NMNL structure for the  $\epsilon_{skh}$ 's, the dependence of choices over a set of  $k = 1, \dots, N_s$  alternative clusters within tract  $s$  may be described in terms of a parameter

---

<sup>6</sup>See Ben-Akiva and Lerman (1985) for a similar approach to decomposing the universe of choices into smaller choice sets.

$\varsigma$ ,  $\varsigma \in [0, 1]$ , which denotes the degree of similarity as reflected in the unobserved component of the evaluation of alternative clusters within each tract. If  $\varsigma = 0$ , then the model implies that alternatives within each tract are independent and the multinomial logit model applies within each tract as well. In that case, the choice probabilities are given by (5). If  $\varsigma \rightarrow 1$ , the other extreme holds and alternatives within a tract are perceived as identical. The model then implies that the choice is made in terms of the maximum value of the utility function.<sup>7</sup>

The first stage of the standard estimation strategy for the NMNL model is to estimate the cluster choice model conditional on census tract choice using MNL. Let  $\mathcal{T}_s$  denote the tract-specific *inclusive* value associated with the choice process over  $N_s$  neighborhood clusters within tract  $s$  by individual  $h$  :

$$\mathcal{T}_s = \ln \left( \sum_{k=1}^{N_s} \frac{1}{N_s} \exp \left[ \frac{1}{1-\varsigma} \left( \beta \Pi_y(y_{n(h)}) + \gamma \Pi_z(z_{n(h)}) + v_k \right) e^{-\delta p_s} \right] \right). \quad (6)$$

The probability of individual  $h$ 's choosing tract  $s_h$  from her choice set,  $s = 1, \dots, S_h$ , becomes:

$$P_{s_h} = \frac{\exp [(1-\varsigma)\ln N_{s_h} + \zeta_h g_{s_h} + (1-\varsigma)\mathcal{T}_{s_h}]}{\sum_{j=1}^{S_h} \exp [(1-\varsigma)\ln N_j + \zeta_h g_j + (1-\varsigma)\mathcal{T}_j]}. \quad (7)$$

As part of this first stage one estimates the inclusive value based on the results of the cluster choice model and substitutes this into the tract choice equation (7). The second stage involves the estimation of the tract choice equation using MNL. A particular problem is that, due to the nature of the NAHS data, we do not observe multiple clusters in a given census tract. Thus we need to devise an alternative estimation procedure for the inclusive value. In doing so, we take a lead from McFadden (1978), p.91–95, regarding aggregation and the treatment of similarities.

The critical step in our procedure relies on the observation that, by the law of large numbers, when the number of neighborhood clusters  $N_s \rightarrow \infty$ , the RHS of (6) may be written in terms of the joint moment generating function of the multivariate distribution of cluster-specific variables

---

<sup>7</sup>This is a special case of the generalized extreme value distribution [ McFadden (1978; 1981; 1984) ]. The NMNL structure may be conveniently described in terms of the so-called generating function

$$H(e^{V_{s1h}}, \dots, e^{V_{skh}}, \dots, e^{V_{sNsh}}) = \left[ \sum_{k=1}^{N_s} e^{\frac{1}{1-\varsigma} V_{skh}} \right]^{1-\varsigma}.$$

This leads, in turn, to a concise description of the overall indirect utility function, as the optimum value of utility associated with the discrete decision problem [ McFadden (1978), Theorem 1, Corollary, p. 538 ], which in our case encompasses the continuous part, as well. We return to this further below.

$(y_{n(h)}, z_{n(h)})$  within each tract.<sup>8</sup> To see this, assume that  $z$  is a scalar and  $(y, z)$  are joint normal  $N(m_y, m_z, \sigma_y^2, \sigma_z^2, \vartheta)$ , and the cluster-specific random effect  $v_k$  has distribution  $N(0, \sigma_v^2)$ . From the definition of the moment generating function  $MGF(B, \Gamma) = \mathcal{E}_{(y,z)}\{\exp[B y + \Gamma z]\}$  it follows that  $\mathcal{T}_s$  may be written in terms of the parameters of the utility function, of the tract-specific price, and of means, variances and covariances  $(m_y, m_z, \sigma_y^2, \sigma_z^2, \vartheta)$  as follows:

$$\begin{aligned} \mathcal{T}_s &= \ell n \left[ \mathcal{E} \left\{ \exp \left[ \frac{1}{1-\varsigma} \left( \beta \Pi_y y_{n(h)} + \gamma \Pi_z z_{n(h)} + v_k \right) e^{-\delta p_s} \right] \right\} \right] \\ &= B m_y + \Gamma m_z + \frac{1}{2} (B^2 \sigma_y^2 + \Gamma^2 \sigma_z^2 + 2\vartheta B \Gamma \sigma_y \sigma_z) + \frac{1}{2} \left( \frac{1}{1-\varsigma} e^{-\delta p_s} \right)^2 \sigma_v^2, \end{aligned} \quad (8)$$

where the auxiliary variables  $B$  and  $\Gamma$  are defined as functions of parameters and of the price,  $B \equiv \frac{1}{1-\varsigma} \beta e^{-\delta p_s}$ ;  $\Gamma \equiv \frac{1}{1-\varsigma} \gamma e^{-\delta p_s}$ . An assumption of “spatial uniformity” (which we think is an appropriate assumption in view of the way tract boundaries are drawn in the U.S.) allows us to identify the mean of each cluster-specific mean with the respective tract-specific mean. Also, we have information about the distributions of a number of tract-level variables including individual and dwelling characteristics that we use to approximate  $\sigma_y^2$  and  $\sigma_z^2$ . There is also some limited information about the joint distribution of some of the tract-level variables that we can use to approximate  $\vartheta$ . Alternatively,  $\mathcal{T}_s$  may be computed non-parametrically from the frequency distributions available in STF3. Both the exact approach using the the moment generating function for a multivariate normal distribution of contextual variables within each tract and the semi-parametric approximation using reported frequencies constitute innovative uses of these data. Our approach subsumes the choice of clusters into the choice over census tracts.

To summarize, our strategy for the neighborhood choice model is to estimate the census tract choice equation (7), with the inclusive value for each of the tracts in an individual’s opportunity set,  $(\mathcal{T}_1, \dots, \mathcal{T}_{S_h})$ , being given by (6) (or by (8), as appropriate). We note that this uses the frequency distributions of the tract-level variables to approximate the inclusive value associated with considering alternative neighborhood clusters within each tract and therefore subsumes the neighborhood choice model within tract choice. We may concisely describe the value to the individual from the

---

<sup>8</sup>See Quigley (1985), who works with a mean-variance approximation to the inclusive value. Our approach exploits better the analytical setting of the problem.

choice process, that is the overall unconditional expected indirect utility function, as:

$$\mathcal{E} [\Omega_{skh}] = \sum_{s=1}^{S_h} \left[ \zeta_h g_s + \left( \alpha + \frac{\mu}{\delta} + \mu p_s + \xi z_h + \delta I_h \right) e^{-\delta p_s} + \mathcal{T}_s \right] P_s, \quad (9)$$

where  $\mathcal{T}_s$  is given from (6) (or by (8), as appropriate) and  $P_s$  from (7), respectively. We underscore that the definition of the inclusive value above subsumes the random component associated with the evaluation of each cluster,  $v_k e^{-\delta p_s}$ .

Conditional on choice of tract according to (7), the cluster choice equation still holds. It implies, in particular, that different households that choose to locate in the same cluster have also chosen to locate in the same tract (by definition). They have done so after having obtained the same evaluation for the cluster they have chosen, which is represented by  $v_k e^{-\delta p_s}$ . Unfortunately, this effect is not identifiable and we proceed to ignore it in the discrete choice part of the problem.

The next problem we confront is how to estimate the tract choice equation given that we have no information about which neighborhoods (either at the level of clusters or of tracts) households have considered before they decided to locate where we observe them. As a first approach, we have implemented an overlooked [ but see Quigley (1985) and Blackley and Ondrich (1988), for two exceptions ] suggestion of McFadden (1978), that the discrete choice model may be estimated by generating a random sample of alternatives from the full choice set (which may be unobserved). Consistency holds provided that if an alternative is included in the assigned set, then it has the logical possibility of being an observed choice from that set [ McFadden, *op. cit.*, 88–89 ].

Our data allow us to generate a random sample of alternative census tracts from among all those comprising the metropolitan area. For each observation in the cluster subsample of the public NAHS data, we identify (using *confidential* U.S. Census data) the tract where it belongs and choose randomly ten tracts, from among the universe of tracts in the respective metropolitan area, which we define as the relevant housing market. Table 3 juxtaposes summary statistics for the actual census tracts in which our AHS observations lie with those for the ten randomly selected tracts. Generally, the sample statistics are similar for the two groups. We estimate a discrete choice model, Equ. (7), where individual  $h$  chooses the tract where he resides out of a set of  $S_h = 11$  tracts that have been chosen randomly from the metropolitan area. We use as regressors tract-level characteristics, on their own and also interacted with individual characteristics. We also include individual variables interacted with quantiles of the distributions of tract-level variables to proxy

for the inclusive value.

### 3.3 Continuous Choice

The discrete and continuous choices modelled here cannot be stochastically specified as independent of one another. Both contribute to the estimation of the behavioral parameters  $(\alpha, \beta, \gamma, \delta, \zeta, \xi, \mu)$  and of the stochastic structure  $\sigma, \sigma_v^2, \sigma_\eta^2$ . If we were to ignore the choice of the neighborhood in estimating the demand model, unobservable elements of the choice problem will affect the consistency of the estimates. Correcting for sample selection bias associated with neighborhood choice is, in and of itself, an important element of the impact of neighborhood effects.

The behavioral model allows for indirect utility, conditional on neighborhood choice, and therefore housing demand as well, to reflect endogenous and contextual neighborhood effects. In specifying such an indirect utility function, we are careful to allow for dependence that is not eliminated by differentiation with respect to price and to income when Roy's identity is applied to obtain housing demand. As Dubin and McFadden (1984) note, the additive shock  $\eta_h$  that is assumed to be present in the demand equation (3) and may be correlated with the unobserved components of utility,  $\epsilon_{skh}$ , must be multiplicatively separable from a function of the housing price in the conditional indirect utility function. This requirement has guided our selection of functional forms and the parameterization of the stochastic structure.

The mean of  $\epsilon_{skh}$ , conditional on alternative  $s_h$  being chosen according to the MNL model, is given by:

$$E[\epsilon_s | s = s_h] = -\frac{\lambda\sqrt{3}}{\pi} \ln P_{s_h}, \quad (10)$$

$$E[\epsilon_s | s \neq s_h] = \frac{\lambda\sqrt{3}}{\pi} \frac{P_s}{1 - P_s} \ln P_s, \quad s = 1, \dots, S_h, s \neq s_h, \quad (11)$$

where  $P_s$  is the probability of choosing tract  $s$ , given by (7) and the distribution of  $\epsilon_s$  is given by

$$\text{Prob}\{\epsilon_s \leq \varepsilon\} = e^{-e^{-\frac{\varepsilon}{\lambda\sqrt{3}} - \tilde{v}}}, \quad (12)$$

and has unconditional mean zero and unconditional variance equal to  $\frac{\lambda^2}{2}$ , and  $\tilde{v} = -\int_0^\infty e^{-g} \ln g dg = .577\dots$ , is Euler's constant. These formulas must be modified if the nested logit (NMNL) model is assumed. See Dubin (1985).

The unconditional mean and variance of the stochastic shock in the continuous equation for housing demand,  $\eta_h$ , are 0 and  $\sigma_\eta^2$ , respectively. The distribution of  $\eta_h$ , conditional on  $(\epsilon_1, \dots, \epsilon_{S_h})$ , has mean  $(\sqrt{2} \frac{\sigma_\eta}{\lambda}) \sum_{m=1}^{S_h} R_m \epsilon_m$ , and variance  $\sigma^2(1 - \sum_{m=1}^{N_s} R_m^2)$ , where  $R_m$  is the correlation coefficient of  $\eta_h$  and  $\epsilon_m$ , and  $\sum_{m=1}^{S_h} R_m = 0$ , and  $\sum_{m=1}^{S_h} R_m^2 < 1$ .

The mean of  $\eta_h$ , conditional on choice  $s_h$  is (for the case of the MNL model)

$$E[\eta_h | s = s_h] = \frac{\sigma_\eta \sqrt{6}}{\pi} \left[ \sum_{s=1}^{S_h} R_s \frac{P_s}{1 - P_s} \ln P_s - R_{s_h} \frac{\ln P_{s_h}}{1 - P_{s_h}} \right]. \quad (13)$$

where  $R_{s_h}$  is the correlation coefficient of  $\eta_h$  and  $\epsilon_{s_h kh}$ . Therefore, consideration of dependence between the random shock determining tract choice and the individual shock affecting housing demand,  $(\epsilon_{s_h kh}, \eta_h)$ , introduces the correlation coefficients  $(R_1, \dots, R_{S_h})$ , between those shocks as additional unknown parameters to be estimated. The conditional second moments for the entire family of generalized extreme value distributions exist in closed form and are given in Dubin (1985), Appendix A. Dubin's formulas can be applied to the case of the nested logit model.<sup>9</sup> The above formulas are indicative of the way in which sample selection bias correction works for the case of errors that are extreme-value distributed.

### 3.4 A Model of Housing Demand, Neighborhood Effects and Neighborhood Choice

We now incorporate the model of neighborhood selection into our model of housing demand with neighborhood effects. We correct for sample selection bias in the standard fashion by writing  $\eta_h = E[\eta_h | s = s_h] + \psi_h$ , where  $\psi_h$  is white noise. We thus have:

$$y_{skh} = \alpha + \mu p_s + \xi z_h + \delta I_h + \beta \overline{y_{n(h)}} + \gamma \overline{z_{n(h)}} + v_k + E[\eta_h | s = s_h] + \psi_h, \quad (14)$$

where the explicit form of the sample selection bias correction term is given in equation (13), and endogenous and contextual effects have been specified in terms of the mean of neighbors' housing demand and characteristics, that is,  $\overline{y_{n(h)}}$  and  $\overline{z_{n(h)}}$ , respectively. Thus, we first estimate the neighborhood choice equation (7) and use the results to estimate the sample selection bias

---

<sup>9</sup>However, Dubin suggests it is better to work directly with the assumption that  $\eta$  has a linear conditional expectation in the space of the "induced" independent extreme value random variables that generate the conditional branch probabilities [ *ibid.*, p. 253 ]. We note that this is important in the context of a suggestion by Moffitt (2001) that an additional route to identification of neighborhood effects may be afforded through the variance-covariance structure. This clearly deserves further attention in future work.



correction terms included in (14). This allows us to obtain consistent estimates of the parameters in the housing demand equation.

Rapaport (1997) estimates a model of housing demand and community choice, without the neighborhood effects, and finds that accounting for community choice makes a big difference in her estimates of the parameters in the demand equation.

### 3.5 Identification of Neighborhood Effects in Housing Demand with Self Selection

As is clear from the housing demand equation, the endogenous social effect  $\overline{y_{n(h)}}$  is correlated with the error term since it includes the unobserved cluster effect,  $v_k$ . Thus we must instrument for  $\overline{y_{n(h)}}$ . As Manski (1993) shows, valid instruments can be difficult to find. Our model of neighborhood choice and housing demand results in a natural set of instruments. As Brock and Durlauf (2001) show, identification of neighborhood effects is enabled by selection bias correction, precisely because in “order to achieve identification, one needs an individual control whose neighborhood average is not a contextual effect” [ *ibid.*, p. 41 ]. This is exactly what selection bias correction introduces into the estimation problem. Using the housing demand equations for all members of a cluster as a system to solve for  $\overline{y_{n(h)}}$ , the cluster means of the sample selection terms arise naturally as identifying instruments:

$$\overline{\ln y_{n(h)}} = \pi_0 + \pi_1 \ln p_\kappa + \pi_2 \overline{z_{n(h)}} + \pi_3 \overline{E[\eta_h | s = s_h]} + \eta_{n(h)}, \quad (15)$$

where  $\overline{E[\eta_h | s = s_h]}$  denotes the means of the selection correction ( “Heckman” ) terms for individual  $h$ ’s neighbors, given by (13),  $\eta_{n(h)}$  is the unobserved error term, and the neighbors’ mean income  $\overline{I_{n(h)}}$  is included in  $\overline{z_{n(h)}}$  for brevity.

Another source of instruments comes from placing our model in the context of the hedonic theory of housing markets. In particular, we include the means of the structural characteristics of neighbors’ dwelling units as instruments. While one might question the validity of the own structural characteristics as instruments, it should be noted that our instruments are the means of one’s neighbors which do not include one’s own values. We can test for the validity of these instruments using an over-identification test. Ioannides and Zabel (2000) do so and do not reject the exogeneity of these instruments. Note that this test requires that one assume that at least one

of the instruments is exogenous. This condition is satisfied in this paper. That is, the means of the neighbors' Heckman terms are exogenous by construction, and therefore, we can test for the exogeneity of all the neighbors' structural characteristics.

In contrast to our model of neighborhood choice when objects of choice are census tracts and contextual effects are appropriately described in terms of the distributions of characteristics within census tracts, our demand model reflects the assumption that each individual knows her neighbors, and demand is a spatial autoregression with *correlated* cluster specific effects. This can be handled, in principle, but it is not a feature of standard packages. If, on the other hand, we were to assume that the endogenous effects are through the expected consumption of one's neighbors (relative to some information set that needs to be specified), then the individual cluster-specific effects are independent across individuals, and the model is therefore easier to estimate.

The contextual effects included in the part of the indirect utility function (1) that affect housing demand comprise the socioeconomic characteristics of one's neighbors. It is possible that other actions of neighbors, such as those associated with *social capital* enter housing demand. Such considerations lead us to recognize that social capital itself may be endogenous and therefore reflects the characteristics of those who decide to locate in different neighborhoods. New issues of identification do arise, in which case the treatment of Durlauf (2001) is particularly apt.

## 4 Estimation

In this section, we estimate the model of neighborhood choice and housing demand, represented by equations (7) and (14). We first discuss two auxiliary estimation tasks that are necessary for estimating the main model. Subsection 4.1 presents the construction of a house price index that is appropriate for our model. Subsection 4.2 follows with a description of our estimates of permanent income. We then turn to the estimation of the model which we carry out in two steps: first, we estimate the neighborhood choice model; second, we use the results of the neighborhood choice model to correct for selection bias and estimate the housing demand equation.

## 4.1 A House Price Index

Estimation of housing demand requires knowledge of the price of housing. The definition of a house price index must be appropriate to the relevant housing market and to the definition of the good and therefore is an important concern of housing research. We have modelled housing demand in terms of a continuous scalar quantity that represents the flow of housing services from both the structure and the neighborhood in which the unit is located. The price of housing is thus the price for a unit of housing services, which is unobserved. We invoke a standard assumption in the housing literature and assume that the market for housing is the MSA. Thus, there is one price per MSA [ Zabel (2002) ].

Recall that the clusters data in our sample come from 100 different MSAs for the years 1985, 1989, and 1993. We were unable to find a price index that covered all 100 MSAs for these three years. To underscore this difficulty, consider for example the Case-Shiller index [ Case and Shiller (1987) ]. It is available for 90 MSAs, but the use of this index results in a loss of about one third of the data since not all MSAs covered by the Case-Shiller index are among the 100 MSAs in our data set. Also, the Case-Shiller index would have to be made comparable across MSAs, which would require a technique like the one we implement below. Therefore, we need to estimate an MSA-specific price index.

A number of methods have been used to obtain the price of housing services. One common approach is to estimate a hedonic house price function,  $P(q_i, x_s; y_{n(h)}, z_{n(h)})$ , as a function of the corresponding vector of structural characteristics of dwelling units,  $q_i$ , of neighborhood attributes,  $x_s$ , and of neighbors' consumption and socioeconomic characteristics,  $(y_{n(h)}, z_{n(h)})$ . In a recent paper, Sieg *et al.* (2001) classify price indices that are based on hedonic regressions into two types: one involves statistical composite commodities and computes the price of a "standard" dwelling unit across markets; the other involves a market-specific scaling factor relative to the price of an "average quality" dwelling unit in the market.<sup>10</sup> Our approach allows for the value of an average quality house to be calculated for different geographical areas that can be used to generate an index of relative house prices. That is, we estimate a hedonic house price model, according to Equ. (16) below, and use the estimated intercept as a house price index.

---

<sup>10</sup>Or put differently, the former holds quantities constant, like the consumer price index, and the latter holds prices constant like the GDP deflator.

Our house price hedonic function involves expressing the log of the reported values of a dwelling unit  $i$  in census tract  $s$  in MSA  $j$  at time  $t$ ,  $\ln W_{isjt}$ , as a function of  $q_{isjt}$ , a vector of structural characteristics of a unit, and of  $x_{st}$ , a vector of neighborhood characteristics that are constant within each cluster and also over a broader geographic area that can account for factors like school quality and safety. The public version of non-neighborhood clusters subsample of the NAHS data does *not* include any such cluster characteristics. We make up for this by using census tract characteristics to proxy for all levels of neighborhood quality. This is made possible by our access to confidential information that allows us to identify the tracts where the individual observations of the NAHS lie. Our use of the non-cluster data affords two advantages. One is that they make up approximately 90% of the NAHS data thus giving us a much larger data set. Second, the prices thus obtained come from a different data set than the one used to estimate the housing demand equation.

We estimate the following house price hedonic model

$$\ln W_{isjt} = a_0 + \sum_{j=1}^{J-1} a_{0jt} MSA_{ijt} + a_1 q_{isjt} + a_2 g_{st} + u_{isjt}, \quad (16)$$

$$i = 1, \dots, N_j, \quad s = 1, \dots, S_j, \quad j = 1, \dots, J, \quad t = 1985, 1989, 1993.$$

The structural characteristics of a unit include the age of the unit and its square, the number of full baths, of bedrooms, and of total rooms, whether or not there is a garage and a number of additional structural quality variables (such as whether the enumerator saw cracks on walls and ceilings, broken pipes, etc.). The neighborhood characteristics include a dummy variable that indicates whether or not the unit lies in the central city of the MSA, the property tax rate, and tract-level variables that include median household income, the percent over 25 years of age who graduated from high school, and the percent of the tract population who are nonwhite.

There are 140 potential MSAs but we require that there are at least ten observations in an MSA in a given year for observations in that MSA to be included in the regression. This leaves 8603, 10083, and 8283 observations for each of 1985, 1989, and 1993, respectively. As we clarify below we estimate the housing demand equation using the pooled data. In order for the price index to be comparable across years, we estimate the house price hedonic using the pooled data. The adjusted  $R^2$  is 0.654. Given that we define each MSA in each year as a separate housing market, it would make sense to allow the coefficients to vary across time and space. This would lead to a very large

number of parameters in the model and very few degrees of freedom in some of the MSAs. Instead, we restrict the coefficients on all variables but the intercept to be constant across MSAs. This is consistent with Mills and Simenauer (1996) who find very little variation in coefficients over time and space when they estimate a house price hedonic using national data from the House Financing Transaction Database, collected by the National Association of Realtors over the period 1986-1992. We do allow the coefficients to vary over time (though not across MSAs) but this decreases the estimated standard error of the regression by only 0.05%. So we use the restricted model for our analysis.

The actual regression results are not reported here because they have not yet been released for disclosure purposes. We note that, as a group, the neighborhood proxies are very significant. We use the estimated MSA-specific intercepts to calculate a price index for the MSA where unit  $i$  lies,  $\hat{p}_{jt} = \hat{a}_{0jt}$ . We set the price for Denver in 1985, the excluded MSA, to be 100 and those for the other MSAs to be 100 times the antilog of the corresponding coefficient estimate.

## 4.2 Estimating Permanent Income

It is standard practice to use permanent rather than current income in housing demand equations [ Olsen (1987); Goodman (1988); Henderson and Ioannides (1989) ]. Also, it is reasonable to assume that individuals are better able to predict their neighbors' permanent rather than current income given the larger fluctuations in the latter measure. We define (the log of) permanent income as the predicted value from the following model of (the natural log of) current income:

$$\ln I_{ht} = \theta_0 + \sum_{j=1}^{J-1} \theta_{0jt} MSA_{hjt} + \theta_1 c_{ht} + \nu_{hjt}, \quad (17)$$

where  $c_{ht}$  includes a cubic polynomial in age and years of education, dummy variables that indicate if the owner is married, male, Black, or Hispanic and whether or not the unit lies in the central city of the MSA. We include the same MSA dummy variables that are in the house price hedonic regressions (16) to capture differences in the cost-of-living across MSAs. We use the non-neighborhood clusters subsample of the NAHS to estimate the model of income for the same reasons that led us to use these data to estimate the price index. We estimate a separate regression for each of the three years. The adjusted  $R^2$ 's for these regressions (with the number of observations in parentheses) for 1985, 1989, and 1993 are 0.405 (17,904), 0.395 (23,610), and 0.329 (19,038), respectively. The actual regression

results have not yet been released for disclosure purposes. We then use the parameter estimates from these regressions, along with the data from the cluster subsample to estimate the permanent income variable that is included in the neighborhood choice and housing demand equations.

### 4.3 Estimation of Neighborhood Choice

We estimate the model of neighborhood choice according to Equ. (7). We use individual and tract-level characteristics as explanatory variables. We select a variety of variables in order to capture possibly complicated effects contributing to the attractiveness of different neighborhoods and their amenities, especially as they pertain to human capital accumulation. We have been influenced in the selection of explanatory variables by such previous studies as Borjas (1998), Nechyba and Strauss (1998), Rapaport (1997), and Quigley (1985). Overman (1999) and Quigley (1985) are the only other studies in the literature that we are aware of that use two different sources of contextual information that are hierarchically related to one another.

We estimate a number of regression models but report only two. The first model is a benchmark model that contains only tract specific variables. From column 1 of Table 4, we see that a higher percentage of ownership in the tract, of residents commuting less than twenty minutes, and of nonwhites in the tract, and a higher median age of dwellings increase the likelihood of choosing a tract. On the other hand, a higher poverty rate, unemployment rate, vacancy rate, median income, median rent, and median age of tract residents decrease the likelihood of choosing a tract. This regression has a pseudo  $R^2$  of 0.0734.

While these results are interesting in their own right, it is clear that preferences for such tract characteristics as those discussed above will vary across demographic groups. Thus, to control for demographic characteristics, we interact the tract-level variables with individual-level variables. These results appear in column 2 of Table 4. First, we create three dummy variables that indicate if the individual is in the first quartile, the middle two quartiles, and the fourth quartile of the tract-specific distribution of income as reported in the STF3 data. We interact these dummy variables with median income, median rent, the vacancy rate, and the homeownership rate. We find that the valuation of median tract income is increasing in individual income. This will actually increase the likelihood of tract choice for those in the upper income quartile. We find that the valuation

of vacancy rates declines with income though it increases in an absolute sense. On the other hand while individuals positively value homeownership rates, this valuation declines with income.

We interact the tract-level race variables with the dummy variables that indicate whether the individual is white or nonwhite. For whites, an increase in the percent nonwhite will decrease the likelihood of tract choice but there is no additional effect if there is at least 50% nonwhites in the tract. For nonwhites, an increase in the percent nonwhite will increase the likelihood of tract choice and there is an additional positive effect if there is at least 50% nonwhites in the tract.

We generate three dummy variables that indicate if the individual has a college degree, at most a high school degree, and no high school degree and interact them with the percent graduated from high school in the tract (25 years of age and older). Those with no high school degree (college degree) are less (more) likely to choose a tract if the percentage of high school graduates increases.

We generate three dummy variables that indicate if the individual is in the first quartile, the middle two quartiles and the fourth quartile of the sample distribution of age. We interact these variables with the median age in the census tract. We find that the impact of an increase in the median age increase with age such that those in the first three quartiles will be less likely to choose the tract while the oldest individuals will be more likely to choose the tract.

We generate three dummy variables that indicate if the individual is in the first quartile, the middle two quartiles and the fourth quartile of the sample distribution of number of persons in the household. We interact these variables with the median number of bedrooms in the census tract. We find that the impact of an increase in the median number of bedrooms will decrease the likelihood that the smallest sized households will choose the tract while there is no significant impact for larger sized households.

These results strongly suggest that individuals like to live with others like themselves. This can perpetuate income inequality since those with the best opportunities at economic success will cluster together.

Next we add tract-level variables interacted with price, which is MSA-specific, and these variables are significant as a group. The addition of these variables is motivated by the structure of the neighborhood choice model.

Finally we include variables that characterize the distributions of the tract-level characteristics.

These terms capture the variation in clusters in each tract and proxy for the inclusive value in the tract choice model. We include tract-level variables that give the 25th and 75th percentile of income, rent and age. We also include variables that indicate the percentage of residents with 8 or less years of education, 9–11 years, 12 years, 13–15 years, and 16 or more years of education. We include a set of variables that give the percent of housing units that are less than two years of age, between 2–5, 5–10, 10–20, 20–30, 30–40, and more than forty years of age. We also include variables that indicate the percentage of units with 1, 2, 3, 4, or 5 or more bedrooms. These variables (interacted with individual-level variables) are significant as a whole and each group of variables is significant as well. This is the model we use to obtain the predicted probabilities that are necessary to compute the sample selection bias correction terms used in the continuous estimation model below. We are aware of the fact that the selection correction terms that we compute are obtained from a discrete choice model with extreme value disturbances but are used with a continuous choice model that is predicated on normal disturbances. However, the import of the Brock–Durlauf theory [ Brock and Durlauf (2001) ] is that it is crucial to have appropriate instruments, in order to be able to estimate the neighborhood effects model, even if they may be semi-parametric or even non-parametric.<sup>11</sup> It is thus behooves us to include complicated interacted variables that allows us to mimic the dependence implied by the definition of the inclusive value above. When a full complement of 87 explanatory variables are included, the pseudo  $R^2$  rises to 0.1889, and the additional variables included over those in the estimation model reported are statistically significant with respect to the likelihood ratio test.

#### 4.4 Estimation of Housing Demand with Neighborhood Effects

The estimation of housing demand along the lines of Equ. (14) is fairly straightforward except for the presence of endogenous neighborhood effects. As discussed in Section 3, we correct for (non-random) neighborhood choice by including an estimate of the expectation of the disturbance term conditional on census tract choice. This results in an additional eleven selection (Heckman) correction terms, one for each of the tract choices in the neighborhood choice model. The addition of these terms in the housing demand equation induces heteroscedasticity. Therefore, we report

---

<sup>11</sup>We thank Steven Durlauf for this point.



heteroscedasticity-robust standard errors. The dependent variable is the log of housing demand. The regressors include the log of house price, the log of permanent income, the number of persons in the household, and dummy variables that indicate if the owner has graduated from high school, is married, is white, and moved in the last five years. The parameter estimates for these dummy variables are suppressed for reasons of confidentiality. Generally, the results for these variables are comparable to the ones in Ioannides and Zabel (2000).

Table 5 reports the estimation results for the housing demand equation. The first set of results pertains to an OLS regression with one observation for each cluster and no neighborhood effects. Therefore, it can serve as a benchmark representing conventional studies of housing demand with our data. The eleven Heckman selection correction terms are statistically significant as a group at the 1% level. Their inclusion raises the adjusted  $R^2$  from 0.1780 to 0.2101, and most of them are significant on their own, too. There is a noteworthy difference in the estimated parameters from the case with no selection correction. The estimated price elasticity is absolutely smaller in the model with selection correction than in the case without,  $-0.175$  versus  $-0.2222$ . The five demographic variables (other than income) continue to be significant as a group when the Heckman terms are included in the model.

Next, we estimate the housing demand equation with neighborhood effects using random effects, where we include the actual mean of neighbors' housing demand among neighbors as a regressor. The associated endogenous social effect is very high, 0.8878, and very significant. We test for the endogeneity of the mean of neighbors' housing demand using a Hausman test and reject the null hypothesis that it is exogenous (p-value  $< 0.0001$ ).

To correct for the endogeneity bias, we first estimate the reduced form equation for the mean of neighbors' housing demand, according to Equ. (15), using in addition to the variables in the RHS of (15) the means of the structural characteristics of neighboring units. One of the important contributions of this paper over Ioannides and Zabel (2000) is the ability to identify the model using the neighborhood means of the Heckman terms. As shown by Brock and Durlauf (2001), these are valid instruments. These terms are very significant in the first stage regression. In our previous paper, we had to assume the exogeneity of one of the neighborhood mean structural terms that we used to identify our model. Here, the presence of the Heckman terms allows us to test

for the exogeneity of all the neighborhood mean structural terms using the over-identification test. The result is that the exogeneity of these variables is not rejected at the 1% significance level. This provides validation for the use of the neighborhood means of the structural characteristics to identify the housing demand model with interactions in Ioannides and Zabel (2000).

We use the results from the first stage regression to consistently estimate the model of housing demand. These results are given in column 3, Table 5. The estimated price and income elasticities,  $-.3019$  and  $.1207$ , respectively, are statistically very significant. While the price elasticity appears to be reasonable, that of income is surprisingly low. However, all studies that have used this particular data set consistently report low estimates for income elasticities [ *c.f* Goodman (1988) ]. The parameter estimates for the own socioeconomic characteristics cannot be reported here, for reasons of confidentiality, but they are not statistically significant as a group. The endogenous neighborhood effect,  $0.6244$ , is large and highly significant. The contextual effects associated with the neighborhood cluster are not individually significant, except for income, which is  $.2600$  and very significant, but are significant as a group. In contrast to our earlier results in Ioannides and Zabel (2000), here we include as regressors the Heckman selection correction terms but they are not significant as a group. When the means of the structural characteristics of neighboring units are not used as instruments, the estimate (from regressions that we did carry out but do not report here) of the endogenous neighborhood effect,  $0.8175$ , is again very high. Also, just as in Ioannides and Zabel (2000), the correction of the standard errors because of the use of predicted values appears to be negligible.

The selection correction terms are significant in the regressions without neighborhood effects but not in those where such effects are included. Clearly, these terms pick up some of the unobserved components due to neighborhood choice that are part of the unobservables in the model without neighborhood effects. This suggests that the neighborhood effects we introduce in the model capture the unobserved individual factors that affect neighborhood choice.

To further probe the possible model uncertainty that may be affecting our estimates, we have also estimated the demand equation by using the structural characteristics of just one neighboring unit as the additional set of instruments instead of the means of all neighboring units. These are the fourth set of results reported on Table 5. We see now that the endogenous social effect is  $.2101$ ,

still very significant but much smaller, and the contextual effect associated with mean neighbors' income is .6088 and thus much larger, than the respective ones when the means of all neighboring units's structural characteristics are used as instruments.

The regressions we report in Columns 3 and 4 of Table 5 include random cluster-specific effects, which explain a large part of the total variance of the regression. Endogenous effects are smaller when we allow for cluster-specific random effects relative to when we do not. Therefore, the fact that different individuals are in the same neighborhood is an important part of the story, that is, even after we have corrected for neighborhood choice.

## 5 Concluding Remarks

Neighborhood choice has never before been examined at this level of detail. Here it does double duty: it is interesting in its own right and serves to correct for selection, when individuals are seen choosing their neighborhood effects via their choice of a dwelling to own and occupy. Generally, socioeconomic characteristics of neighborhoods appear to be important determinants of neighborhood choice. In particular, we find that individuals prefer to live with others like themselves. This result seems to be consistent with the observed sorting of individuals in the U.S. by income, race, and education.

Our estimates of the housing demand equation are quite similar to those of Ioannides and Zabel (2000) and do confirm that neighborhood effects are important. Neighborhood choice appears to be associated with the following differences, when it is accounted for relative to when it is not, as in our earlier results. First, there is a smaller endogenous social effect: the elasticity of housing demand with respect to mean neighbors' demand is 0.6244 instead of 0.660 in the earlier paper. Second, there is a larger contextual effect: the elasticity of housing demand with respect to mean neighbors' income is 0.2600 instead of 0.175 in the earlier paper.

All these parameters are estimated with great precision. Recall that our procedure for accounting for individuals' opportunity sets assumes that individuals choose the census tract in which we observe them over ten census tracts that we select randomly from within the same metropolitan area. This arguably biases results in favor of making neighborhood choice more significant. The fact that we find few differences with the estimates obtained when neighborhood choice is ignored

is, therefore, quite interesting.

Overall, this research heeds the call by Schelling (1978) on the importance of micro neighborhood interactions and by Glaeser (2000) and others on the need to assess the quantitative significance of urban interactions. Alternative ways for making up for the unobservability of individuals' opportunity sets, such as limiting the full set to tracts with similar characteristics, deserve attention in future research.

## 6 References

- Blackley, Paul and Jan Ondrich (1988), “A Limiting Joint-Choice Model for Discrete and Continuous Housing Characteristics,” *The Review of Economics and Statistics*, 70, 266–274.
- Borjas, George J. (1998), “To Ghetto or Not to Ghetto: Ethnicity and Residential Segregation,” *Journal of Urban Economics*, 44, 228–253.
- Brock, William A., and Steven N. Durlauf (2001) “Interactions-Based Models,” 3297–3380, in Heckman, James J., and Edward Leamer, eds., *Handbook of Econometrics*, Volume 5, North-Holland, Amsterdam.
- Dubin, Jeffrey F. (1985), “Conditional Moments in the Generalized Extreme Value Family,” Appendix B, 217–254, in *Consumer Durables Choice and the Demand for Electricity*, North-Holland, Amsterdam.
- Dubin, Jeffrey F., and Daniel McFadden (1984), “An Econometric Analysis of Residential Electrical Appliance Holdings and Consumption,” *Econometrica*, 52, 345–362.
- de Souza Briggs, Xavier (1997), “Moving Up versus Moving Out: Neighborhood Effects in Housing Mobility Programs,” *Housing Policy Debate*, 8, 195–233.
- Durlauf, Steven N. (2001), “On The Empirics of Social Capital,” University of Wisconsin, mimeo.
- Epple, Dennis, and Holger Sieg (1999), “Estimating Equilibrium Models of Local Jurisdictions,” *Journal of Political Economy*, 107, 4, 645–681.
- Epple, Dennis, Thomas Romer, and Holger Sieg (2001), “Interjurisdictional Sorting and Majority Rule: An Empirical Analysis,” *Econometrica*, 69, 6, 1437–1465.
- Gans, Herbert (1967), *The Levittowners: Ways of Life and Politics in a New Suburban Community*, Columbia University Press, New York.
- Goodman, A.C., (1988), “An Econometric Model of Housing Price, Permanent Income, Tenure Choice, and Housing Demand,” *Journal of Urban Economics*, 23, 237–353.
- Glaeser, Edward L. (2000), “The Future of Urban Research: Non-Market Interactions,” *Brookings-Wharton Papers on Urban Affairs*, 101–138.
- Henderson, J. Vernon, and Yannis M. Ioannides (1989), “Dynamic Aspects of Consumer Decisions in Housing Markets,” *Journal of Urban Economics*, 26, 212–230.
- Ioannides, Yannis M. (2002) “Residential Neighborhood Effects,” *Regional Science and Urban Economics*, 32, 2, 145–165. [PDF at <http://ase.tufts.edu/econ/papers/index.html>].
- Ioannides, Yannis M., and Jeffrey E. Zabel (2000), “Neighborhood Effects and Housing Demand,” presented at the Conference on Empirics of Social Interactions,” Brookings Institution, January 14–15, 2000; *Journal of Applied Econometrics*, forthcoming. [PDF at <http://ase.tufts.edu/econ/papers/index.html>].
- Katz, Laurence, Jeffrey K. Kling and Jeffrey. Liebman (2001), “Moving to Opportunity in Boston: Early Impacts of a Housing Mobility Program,” *Quarterly Journal of Economics*, 116, 2, 607–654.

- Kiel, Katherine A., and Jeffrey E. Zabel (1997) "Evaluating the Usefulness of the American Housing Survey for Creating House Price Indices," *Journal of Real Estate Finance and Economics*, 14, 189-202.
- McFadden, Daniel (1978), "Modelling the Choice of Residential Location," in Karlqvist, Anders, *et al.*, eds., *Spatial Interaction Theory and Planning Models*, North-Holland, Amsterdam, 75-96.
- McFadden, Daniel (1981), "Econometric Models of Probabilistic Choice," in Manski, Charles F., and Daniel McFadden (1981), *Structural Analysis of Discrete Data with Econometric Applications*, MIT Press, 198-272.
- McFadden, Daniel (1984), "Econometric Analysis of Qualitative Response Models," Ch. 24, 1395-1457, in Griliches, Zvi, and Michael D. Intriligator, eds., *Handbook of Econometrics*, Volume II, North-Holland, Amsterdam.
- Manski, Charles F., (1993) "Identification of Endogenous Social Effects: The Reflection Problem," *Review of Economic Studies*, 60, 531-542.
- Nechyba, Thomas, J., and Robert P. Strauss (1997), "Community Choice and Local Public Services: A Discrete Choice Approach," *Regional Science and Urban Economics*, 28, 51-73.
- Neisheim, Lars (2001), "Equilibrium Sorting of Heterogeneous Consumers across Locations: Theory and Empirical Implications," Ph.D. dissertation, Department of Economics, University of Chicago, August.
- Olsen, Edgar O., (1987), "The Demand and Supply of Housing Service: A Critical Survey of the Empirical Literature," 989-1022, in Edwin S. Mills ed., *Handbook of Regional and Urban Economics*, Volume 2, Amsterdam: North-Holland.
- Overman, Henry G. (1999), *Empirical Studies on the Location of Economic Activity and its Consequences*, Ph. Thesis, London School of Economics, September.
- Parzen, Emmanuel (1960), *Modern Probability Theory and Its Applications*, John Wiley and Sons, New York.
- Pollak, Robert A. (1976), "Interdependent Preferences," *American Economic Review*, 66, 309-320.
- Quigley, John M. (1985), "Consumer Choice of Dwelling, Neighborhood and Public Services," *Regional Science and Urban Economics*, 15, 41-63.
- Rapaport, Carol (1997), "Housing Demand and Community Choice: An Empirical Analysis," *Journal of Urban Economics*, 42, 243-260.
- Schelling, Thomas C. (1978), *Micromotives and Macrobehavior*, Norton, New York.
- Sieg, Holger, V. Kerry Smith, H. Spencer Banzhaf, and Randy Walsh (2001), "Interjurisdictional Housing Prices in Locational Equilibrium," Duke University working paper, July.
- Tiebout, Charles M. (1956), "A Pure Theory of Local Public Expenditures," *Journal of Political Economy*, 64, 416-424.
- Zabel, Jeffrey E. (2002), "Continuous Models of Housing Demand," working paper, Tufts University.

**Table 1: SUMMARY STATISTICS:****American Housing Survey, 1985, 1989, 1993 6372 Observations**

Variable	Mean	Standard Deviation
Permanent income (log)	10.23844	.7007072
Household heads with high school education, in cluster (fraction)	.8435527	.1804457
Nonwhite household heads, in cluster (fraction)	.1534013	.2918351
Moved within 5 years, in cluster (fraction)	.3399812	.1973038
Mean age of household heads in cluster	51.51979	7.071606
Owners in cluster (fraction)	.8577508	.1565115
Mean permanent income among neighbors (log)	10.22315	.5171604
Household heads married, in cluster (fraction)	.6312261	.2105387
Mean neighbor household size	2.77464	.6574765
Value of dwelling unit (\$)	121370.3	105472.6
Mean age of dwelling unit (years)	37.03751	16.38772
Mean of full baths, neighbor units	1.640929	.6021015
Mean of number of bedrooms, neighbor units	3.020716	.5686652
Mean of garage, neighbor units	.7934714	.2845211
Mean of number of rooms, neighbor units	6.419021	1.147638
Air conditioning, neighbor units	.7532957	.3023003
Cracks in walls, neighbor units	.0340552	.0759873
Holes in walls, neighbor units	.0061205	.0302858
Major structural defects, neighbor units	.029818	.0700384
Household head schooling (years)	13.26899	3.201469
Size of household	2.761299	1.461693
Log of price	4.416619	.3397629
Price index (Denver MSA=100)	87.80374	30.73653
Age of household head (years)	53.33914	15.89215
Observations in 1985 ( = 1, if in)	.2741682	.4461292
Observations in 1989 ( = 1, if in)	.3066541	.4611407
Observations in 1993 ( = 1, if in)	.4191777	.4934632
Mean predicted housing demand of neighbors	7.018893	.5941519

**Table 2: FREQUENCY OF NEIGHBORHOOD CLUSTER SIZES**

Units in cluster	5	6	7	8	9	10	> 10
Number of clusters	370	534	665	976	1188	1380	1259

**Table 3: SUMMARY STATISTICS: STF3A, 1980, 1990**

	Census Tracts of AHS Data		10 Census Tracts Randomly Chosen, same MSA	
Observations	6372		63720	
Variable	Mean	S.D.	Mean	S.D.
Unemployment rate	.0560639	.0358774	.0753302	.0585049
Unemployment rate $\times$ % non-whites	.0136405	.0388313	.0111751	.0358358
Poverty rate	.0911387	.0855463	.1435074	.132609
Fraction non-whites	.2105101	.2626059	.2821489	.307437
Fraction vacancies	.0562693	.0470073	.0760729	.0633072
Fraction owners	.6830297	.1781918	.5717286	.2350113
Fraction changed hands in last 5 years	.4494394	.1321869	.4902723	.1423698
25-percentile of income (000\$)	25.74783	10.96303	21.43516	11.65585
Median income (000\$)	38.88759	13.92264	34.01683	14.97174
75-percentile of income (000\$)	53.90132	17.69408	48.94414	19.04568
Median age of individuals	31.48832	5.233546	30.47135	5.701576
Median number of bedrooms	2.716573	.5664355	2.487398	.6140009
Median age of dwelling units	27.07454	10.56667	26.98511	11.16377
Fraction complete high school	.7301739	.1154721	.6885016	.1488004
Fraction completed $\leq 8$ years	.0779382	.0667831	.1035867	.0943983
Fraction completed $9 \leq \cdot \leq 11$	.1293679	.0697979	.1512281	.0858681
Fraction completed $13 \leq \cdot \leq 15$	.1996313	.0558615	.1882685	.0597654
25-percentile of monthly rent (\$)	360.8111	103.831	340.7442	109.0499
Median rent (\$)	415.1716	82.46215	398.4197	90.94635
75-percentile of monthly rent (\$)	504.857	94.72476	485.6015	104.6676
White $\times$ Fraction non-white	.1120112	.1485008	.2302386	.2947284
Non-White $\times$ Fraction non-white	.0984989	.2626382	.0519103	.1776417
Dominant race non-white=1, if Percent non-white $> .50$	.1442247	.351345	.213371	.4096907



Table 4:

**MULTINOMIAL LOGIT MODEL FOR CHOICE OF CENSUS TRACT OF RES-  
IDENCE: CENSUS TRACT- AND INDIVIDUAL-LEVEL VARIABLES**

Variable	Restricted Model	Full Model
Median income	-.0043881* (.00201)	
Median income $\times$ income in 1st quartile		-.0370789** (.0045053 )
Median income $\times$ income in 2nd, 3rd quartiles		-.0133117** (.0027746 )
Median income $\times$ income in fourth quartile		.0074424* ( .0029654)
Median rent	-.0010513** (.0002937)	
Median rent $\times$ income in 1st quartile		-.0007832 (.0005963 )
Median rent $\times$ income in 2nd and 3rd quartile		-.0013733** (.0004271)
Median rent $\times$ income in 4th quartile		-.0007843 ( .0005972)
Fraction non-white in tract	.4079781** (.1362475)	
White $\times$ Fraction non-white in tract		-.7515546** ( .1718891 )
Non-white $\times$ Fraction non-white in tract		4.198833** (.2860615)
Dominant race	.2573424** (.087396)	
Dominant race $\times$ hh head white		-.2153893 (.1182571)
Dominant race $\times$ hh head nonwhite		.5873344** (.189959)
Percent commute less than 20 minutes	1.062809** (.1531067)	1.649353** (.2828286)
Percent commute less than 20 minutes $\times$ hh head male		-.6188855 ( .3247698 )

**Table 4: Continued**

Variable	Restricted Model	Full Model
Poverty rate	-2.364136** (.3407552)	-3.528244** (.3755795)
Unemployment rate	-5.795275** (.6992684)	-8.364291** (.7560274)
Fraction completed high school	.1479637 (.213365)	
Fraction completed high school × hh head at most completed high school		-.4033415 (.2585439)
Fraction completed high school × hh head not completed high school		-4.097066** (.3444453)
Fraction completed high school × hh head completed college		3.815756** (.3493936)
Median age indiv in tract	-.0109608** (.0034807)	
Median age of house in tract × hh h age in 1st quartile		-.0393828** (.0082992)
Median age of house in tract × hh h age in 2nd, 3rd quartile		-.0190462** (.0064139)
Median age of house in tract × hh h age in 4rth quartile		.0134401* (.0062834 )
Median age of house in tract ×hh head married		-.0055739 (.0061202 )

**Table 4: Continued**

Variable	Restricted Model	Full Model
Median age of house in tract	.0049918** (.0016002)	
Median number of bedrooms	.0702261 (.0370069)	
Median number of bedrooms × hh size in 1st quartile		-.2183383** (.0554643)
Median number of bedrooms × hh size in 2nd, 3rd quartile		.0464803 (.0834342)
Median number of bedrooms × hh size in 4rth quartile		.0084156 (.077657)
Median number of bedrooms × married		.2294461** (.0628937)
Fraction moved within last 5 years	-.1868096 (.1978472)	
Fraction moved within last 5 years × hh head age in 1st quartile		.7771453* (.3164385)
Fraction moved within last 5 years × hh head age in 2nd, 3rd quartile		-.260743 (.2560377)
Fraction moved within last 5 years × hh head age in 4rth quartile		.3948822 (.3225215)
Fraction owners	1.794969 ** (.1435287)	
Fraction owners × income in 1st quartile		2.430609 ** (.2448994)
Fraction owners × income in 2nd, 3rd quartile		1.88374 ** (.1787906)
Fraction owners × income in 4rth quartile		1.011504 ** (.2218308)

**Table 4: Continued**

Variable	Restricted Model	Full Model
Fraction vacancies	-2.934503** (.4024874)	
Fraction vacancies $\times$ income in 1st quartile		-1.118151** (.6835639)
Fraction vacancies $\times$ income in 2nd, 3rd quartile		-3.440058** (.5881687)
Fraction vacancies $\times$ income in 4rth quartile		-7.924718 ** (1.080726)
Log of tract size	-.0200521 (.0299843)	-.0098041 (.0310739 )
Observations	70092	70092
Log likelihood	-14158	-12796
$\chi^2$ Significance, all variables	.0000	.0000
Pseudo $R^2$	.0734	0.1625

Table 5

## ESTIMATION OF DEMAND EQUATION

(Robust Standard Errors in Parentheses)

Model	OLS	OLS	POOLED	POOLED-alt.instr
Cluster information	No	Yes	Yes	Yes
Dummy =1, if in, in 1989	.1983** (.0527)	.0306* (.0125)	.0975** (.0346)	.2055** (.0336)
Dummy =1, if in, in 1993	.1566** (.0509)	.0057 (.0118)	.0052 (.0312)	.0229 (.0312)
Mean observed demand by neighbors		.8878** (.0147)		
Mean Predicted demand by neighbors			.6244** (.0542)	.2101** (.0379)
Log price	-.1892* (.0847)	-.0823* (.0303)	-.3019** (.0841)	-.5971** (.0793)
Log income	.4091** (.0551)	.0743** (.0163)	.1207** (.0166)	.1458** (.0165)
Household size	.0120 (.0153)	.0113** (.0036)	.0083* (.0041)	.0071 (.0041)
Completed high school	– (–)	– (–)	– (–)	– (–)
Changed hands within last 5 years	– (–)	– (–)	– (–)	– (–)
White	– (–)	– (–)	– (–)	– (–)
Married	– (–)	– (–)	– (–)	– (–)
Mean Log income, neighbors		.0263 (.0412)	.2600** (.0771)	.6088** (.0702)
Mean household size, cluster		-.0156 (.0095)	-.0205 (.0227)	-.0359 (.0229)
Fraction completed high school, cluster		.0098 (.0352)	.0087 (.0852)	.0740 (.0845)

Note: \*, \*\* indicate significant at the 5, 1 per cent significance levels. – indicates suppressed for confidentiality.

**Table 5: CONTINUED**

Fraction changed hands in last 5 years, cluster		-.0307 (.0234)	-.0542 (.0580)	-.1168* (.0586)
Fraction non-white, cluster		-.0085 (.0278)	-.0065 (.0506)	-.1168* (.0586)
Fraction married, cluster		-.0059 (.0381)	-.0989 (.0926)	-.0169 (.0510)
CONSTANT	3.2691** (.4388)	.1678* (.1426)	.2167 (.3526)	.6128 (.3514)
Heckman Correction	Yes	Yes	Yes	Yes
Prob $F$ – test, Heckman terms	0.0121	0.2988	0.0181	0.0365
Prob $F$ – test, own socioeconomics	0.0088	????	0.0803	0.1856
Prob $F$ – test, cluster socioeconomics		????	0.0064	0.0000
Observations	764	6372	6372	6372
Observations per cluster	1	8.3	8.3	8.3
R-SQUARE: WITHIN		0.3722	0.0009	0.0099
R-SQUARE: BETWEEN		0.9960	0.6150	0.4896
R-SQUARE: OVERALL	0.2101	0.6456	0.4430	0.3546
S.D of RE		0	.2907	.2911
S.D. Regression Error	.5451	.0963	.3276	.3331
Per cent Variance due to RE		0	.4404	.4330
Prob. $\chi^2$ , random effects		.0000	.0000	.0000